UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Rotello, C. M., et al. (2018). The Shape of ROC Curves in Shooter Tasks: Implications for Best Practices in Analysis. *Collabra: Psychology,* 4(1): 32. DOI: https://doi.org/10.1525/collabra.171

ORIGINAL RESEARCH REPORT

The Shape of ROC Curves in Shooter Tasks: Implications for Best Practices in Analysis

Caren M. Rotello*, Laura J. Kelly† and Evan Heit†

Four experiments addressed the widely studied issue of the association between racial groups and guns, namely shooter bias, as measured in the first-person shooter task or the weapon identification task, in which participants judge whether a suspect has a weapon or some other item such as a phone (Correll, Park, Judd, & Wittenbrink, 2002; Payne, 2001). Previous studies have employed various analyses that make conflicting, and indeed untested, assumptions about the underlying nature of the data: Analyses of variance and model-based analyses assume linear receiver operating characteristics (ROCs) and signal detection (SDT) analyses assume curved ROCs. The present experiments directly investigated the shape of the ROCs for the weapon identification task, demonstrating that they are curved, and that the majority of previous studies are at risk for inclusion of inappropriate analyses, because they assume linear rather than curved ROCs.

Keywords: Shooter bias; Weapon identification; Signal detection; Reproducibility

One of the key conceptual points in social psychology is that there are implicit attitudes, which are manifest not by direct reports but rather by their effects on judgment and behavior (Greenwald & Banaji, 1995). The concept of, and the existence of, implicit attitudes is a matter of great scientific and societal importance due to the potential impact of harmful stereotypes on everyday interactions as well as institutions connected to the legal system, education, employment, medicine, and so on. In scientific research, the most well-known measure of implicit attitudes is the Implicit Association Test (Greenwald, McGhee, & Schwartz, 1998), which aims to uncover the associations between concepts, e.g., from Caucasian and African-American to good and bad, by recording response times to a choice task. Likewise, other tasks have been employed by researchers to investigate associations between concepts without the requirement of making a direct report. Here, our focus is on shooter tasks, including the *weapon identification task* and the *first-person shooter task.*

Having scientific as well as real-world, even life-or-death, implications, shooter tasks have been used to investigate the association between racial groups and guns (e.g., Correll, Park, Judd, & Wittenbrink, 2002; Greenwald, Oakes, & Hoffman, 2003; Payne, 2001). In a typical study, subjects see pictures of White and Black criminal suspects, paired with either a gun or a non-lethal object such as a cell phone. In the weapon identification version of the task, subjects must decide quickly whether the object that appears after a face prime is a gun or not. In the first-person shooter version of the task, subjects decide whether to shoot the suspect (because he is holding a gun) or to not shoot (because he has a cell phone). These studies have generally reported a negative bias towards Black suspects. However, the nature of the reported shooter bias has varied considerably. In some cases, there is a reported overall response tendency to make a "gun" response (or shoot more often) with Black suspects. In other cases, there is a particular focus on false alarms, i.e., shooting at unarmed suspects, as in Correll et al. (2002). In other studies, the focus is on accuracy, i.e., correctly responding to guns as guns and to non-guns as non-guns. Moreover, in numerous studies, the focus is not so much on the responses but on the response times, e.g., whether subjects are faster to make gun responses to Black suspects than to White suspects.

Considering this variation in how performance in these shooter tasks is measured, it may not be surprising that the findings from studies involving the task have themselves been mixed. A recent meta-analysis (Mekawi & Bresin, 2015) of 42 studies across 16 articles and theses utilizing the first-person shooter paradigm or variant thereof concluded that there is an overall tendency to respond "shoot" or "gun" more often with Black suspects, measured in terms of a particular response bias statistic, *c*, that is derived from signal detection theory (SDT, Macmillan & Creelman, 2005). This bias effect is highly variable across participants and studies. However, there was no statistically significant effect of suspect race on the false alarm rate, nor on decision accuracy, measured with the *d'*

* University of Massachusetts, US

† University of California, Merced, US

Corresponding author: Caren M. Rotello (caren@psych.umass.edu)

measure from SDT. Additionally, Mekawi and Bresin (2015) concluded that there were significant effects of race on response time: participants were faster to shoot at armed Black suspects and slower to avoid shooting at unarmed Black suspects.

The foregoing discussion suggests that there is not a consistent operational definition of shooter bias in the literature, and that these findings may depend on the definition that is used. In many cases, the operational definition is likely determined by the experimental design. For example, a study that requires speeded responses may expect to see effects in error rates, whereas a design that imposes less time pressure may expect to reveal differences in reaction times. If these different dependent variables are not taken into account when forming an analysis plan the use of multiple measures can, of course, increase the risk of a Type I error (Simmons, Nelson, and Simonsohn; 2011; see also Mekawi & Bresin, 2015). We will argue that there is an additional problem in studies of shooter bias. Specifically, studies of the first-person shooter and weapon identification tasks have employed fundamentally incompatible measures of performance, typically within the same study. Put simply, different measures of shooter bias make different statistical assumptions; violating those assumptions can also increase the risk of erroneously concluding that there is a difference in decision accuracy as a function of race (e.g., Swets, 1986; Rotello, Masson, & Verde, 2008).[1] To the extent that one measure is appropriate or well-justified for a particular study, another will be inappropriate and unjustified, for that same study (e.g., Pazzaglia, Dube, & Rotello, 2013; Heit & Rotello, 2014). Moreover, there is little evidence in the extant literature to allow researchers to decide which measure to use. Because our statistical concerns relate to the responses themselves, rather than to response times (which are also widely reported in first-person shooter and weapon identification studies), our focus here will be on measures of response rates. However, we will make some comments about response times in the General Discussion.

To motivate these concerns further, we present our own reckoning (**Table 1**) of dependent measures and analyses in various studies of these tasks. Sixteen of the 45 journal articles and unpublished theses of **Table 1** were included in the Mekawi and Bresin (2015) meta-analysis, which had inclusion criteria of having a simultaneous agent and object presentation (the classic Correll et al., 2002, procedure, or a task with a face and object superimposed on each other), compared responses to White and Black agents along response time or error rate metrics, and had the appropriate reporting to calculate effect sizes. The remaining studies consist of weapon identification (serial face-then-object presentation) tasks, go/no-go shooter tasks, did not provide information to calculate relevant effect sizes or were published after the meta-analysis was conducted. As can be seen, the majority of these studies (29 out of 45, or 64%) report analyses of variance (ANOVAs) on response rates in terms of raw scores (i.e., P("gun"), P("shoot"), or P(error) in each condition). These ANOVAs can be used for various purposes, for example to examine false alarm rates, overall rates of responding, or other terms of interest, such as the interaction between race of the suspect and the presence of a gun on the probability of decision to shoot. Many of these studies (24 out of 45, or 53%) report SDT analyses, employing measures such as *d'* and *c* to examine response accuracy and bias. Finally, some studies employ model-based analyses, using Jacoby's (1991) *process dissociation procedure* (PDP, 11 of 45, or 24% of studies) or Sherman et al.'s (2008) Quad model (4%). The PDP approach is an algebraic method for estimating the contributions of two processes, such as a "controlled" response to the presence or absence of a weapon versus an "automatic" response based on a stereotypical bias; the Quad model is similar but estimates two additional parameters, for potentially overcoming the bias as well as guessing.

The key to answering the question of which analytic approach is most appropriate for the data is to use *receiver operating characteristics* (ROCs, Macmillan & Creelman, 2005). ROC curves are used to plot results from detection studies (see **Figure 1** for examples). The ROCs show the relationship between hit rates and false alarm rates across different levels of confidence or response bias; all points on an ROC reflect the same decision accuracy according to some measure of performance (e.g., *d'*, percent correct, false-alarm rate). The two main possibilities for the relationship between hit and false alarm rates are captured with linear and curved ROCs. It is well-established that ANOVAs on raw response probabilities assume linear ROCs (Dube, Rotello, & Heit, 2010). Likewise, the PDP method assumes linear ROCs (Buchner, Erdfelder, & Vaterrodt-Plünnecke, 1995; Pazzaglia et al., 2013; Rotello et al., 2015). The Quad model also generates linear ROCs (Heit & Rotello, 2014, Rotello, Heit, & Dubé, 2015). In contrast, SDT measures such as *d'* and *c* assume symmetrically curved ROCs (Macmillan & Creelman, 2005). So, for example, if any individual study reports both ANOVAs on raw scores as well as SDT measures, these are incompatible analyses. Both cannot be appropriate for the same study because they make fundamentally different assumptions about the relationship between hit and false alarm rates as a function of bias (see, e.g., Swets, 1986a; Macmillan & Creelman, 2005; Rotello, Masson, & Verde, 2008). Moreover, if the ROCs are curved, then ANOVAs on raw scores and PDP/Quad model analyses are inappropriate. In that case, any of these approaches may lead to false conclusions because the measures confound decision accuracy with response bias. Finally, if the ROCs are linear, then the previously reported SDT analyses are incorrect; and if the ROCs are curved but asymmetric, then different SDT analyses are required.

The studies listed in **Table 1** attempted to do more than simply establish whether there is an effect of race on the identification of a gun or the decision to shoot: They also attempted to discover moderators of those responses, such as participant demographics, time available to respond, and salience of race in the particular context. The conclusions about moderators of shooter bias are also at risk. To give one example from the meta-analysis itself, Mekawi and Bresin (2015) reported that racial diversity of

Table 1: Methods of Analysis of Previously Reported Weapons Identification and First-Person Shooter Tasks.

Citation	ANOVA or Regression on Error Rates	SDT	Modeling	Reaction Times
Akinola & Mendes (2012)[1]	•	•	–	–
Amodio et al. (2004)	•	–	PDP	•
Brewer, Ball, & Ware (2016)	–	•	–	–
Conrey et al. (2005)	–	–	Quad	–
Correll (2008)	•	–	–	•
Correll, Park, Judd, & Wittenbrink (2002)[1]	•	•	–	•
Correll, Park, Judd, & Wittenbrink (2007b)[1]	–	•	–	•
Correll, Park, Judd, Wittenbrink et al. (2007a)	–	•	–	•
Correll, Urland, & Ito (2005)	–	–	–	•
Correll, Wittenbrink, Crawford, & Sadler (2015)	•	•	–	•
Correll, Wittenbrink, Park, Judd, & Goyle (2011)[1]	•	•	–	•
Cox, Devine, Plant, & Schwartz (2014)	•	–	–	•
Fleming, Bandy, & Kimble (2010)	•	•	–	•
Govorun & Payne (2006)	•	–	PDP	–
Greenwald, Oakes, & Hoffman (2003)	–	•	–	–
Harmer (2012)[1]	•	•	–	•
Hunsinger (2011)[1]	–	•	PDP	–
Huntsinger, Sinclair, & Clore (2009)	•	–	PDP	•
Ito et al. (2015)	–	–	PDP	–
James, Klinger, & Vila (2014)	–	–	–	•
James, Vila, & Daratha (2013)	•	–	–	•
Kahn & Davies (2010)	•	•	–	–
Klauer & Voss (2008)	•	–	PDP	•
Kleider & Parrott (2009)	•	–	–	–
Kleider, Parrott, & King (2010)	•	•	–	•
Lambert et al. (2003)	•	–	PDP	–
Ma & Correll (2011)	•	–	–	–
Ma et al. (2013)	•	•	–	•
Mekawi, Bresin, & Hunter (2016)[1]	–	•	–	–
Miller, Ziwlaskowski, & Plant (2012)[1]	•	–	–	–
Musolino (2012)[1]	•	•	PDP	•
Park & Glaser (2011)[1]	–	–	–	•
Park et al. (2008)[1]	–	–	–	•
Payne (2001)	•	–	PDP	•
Payne, Lambert, & Jacoby (2002)	•	–	PDP	•
Plant & Peruche (2005)[1]	•	•	–	–
Plant et al. (2011)[1]	•	•	–	•
Plant, Peruche, & Butz (2005)[1]	•	–	PDP	•
Sadler et al. (2012)[1]	–	•	–	•
Senholzi et al. (2015)	–	–	–	•

(contd.)

Citation	ANOVA or Regression on Error Rates	SDT	Modeling	Reaction Times
Sim et al. (2013)	–	•	–	•
Taylor (2011)[1]	•	•	–	•
Tenhundfeld (2015)	–	•	–	•
Watt (2010)	•	–	–	•
Witt & Brockmole (2012)	–	•	–	•

[1]Included in Mekawi and Bresin (2015) Meta-Analysis.

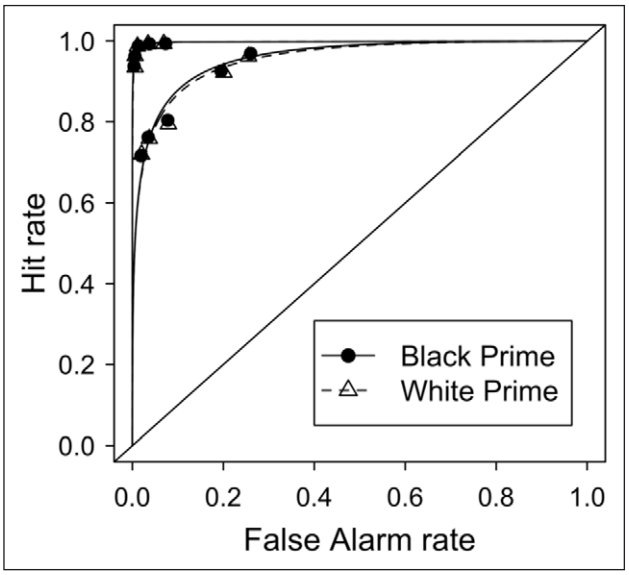

Figure 1: Receiver operating characteristic curves (ROCs) for the data from Experiments 1a (200 ms displays, higher points) and 1b (50 ms displays, lower points). The symbols indicate the observed response rates at each level of decision confidence. The superimposed functions are the best-fitting ROCs for the Black and White prime conditions fit separately.

the city where the study was conducted was a significant moderator of shooter bias in terms of an SDT measure of response bias, but only a marginal moderator in terms of a raw score measure of response tendency (i.e., false alarm rate). So, determining whether local racial diversity moderates the shooter bias depends on which measure is appropriate. Broadly speaking, considering the potential scientific, not to mention societal, importance of weapon identification task and first-person shooter task studies, it is extremely important to check that the correct analyses are being used. To our knowledge, no researcher has published ROC analyses of a weapon identification or shooter task. A priori, we have suggested that because both types of studies are essentially visual detection tasks, the empirical ROCs are likely to be curved (Rotello et al., 2015), as in virtually all perception tasks for which there are ROC data (Dube & Rotello, 2012).

One straightforward way to obtain ROC curves simply requires that participants make confidence ratings along with their gun/non-gun responses. Simple binary response tasks, like the original weapon identification and first-person shooter tasks, require participants to set a single decision criterion; evidence values that exceed the criterion elicit one type of response (e.g., "gun"/"shoot"), and those that fall below that criterion result in the alternative response (e.g., "tool"/"don't shoot"). Confidence ratings can be used to further divide the evidence scale with additional decision criteria, from "sure gun" decisions to "sure tool" responses, for example. In that case, the highest-confidence "gun" responses to guns and tools define the hit and false alarm rate plotted with the left-most point on the ROC; as confidence in "gun" decisions declines, the other points on the ROC are established (see Macmillan & Creelman, 2005, for details). Both binary response and confidence-rating tasks have been used extensively in other research domains to study similar types of classification tasks (e.g., recognition memory: Egan, 1958; Snodgrass & Corwin, 1988; Dube & Rotello, 2012; eyewitness identifications: Mickes, Flowe, & Wixted, 2012; reasoning: Dube, Rotello & Heit, 2010; Heit & Rotello, 2014; medical diagnosis: Swets et al., 1979; weather forecasting: Mason, 1982; and visual and auditory perception: Swets, Tanner, & Birdsall, 1961; see Swets, 1986b, for a partial summary). Using examples from several of these domains, Rotello, Heit, and Dubé (2015) showed that ROC analyses provide a significant advantage over binary-response analyses precisely because they allow identification of the most appropriate test statistic to summarize decision accuracy, namely one that is independent of response bias (see also, Pazzaglia, Dube, & Rotello, 2013; Rotello et al., 2008).

In the present work, we aimed to directly investigate the shape of ROCs in weapon identification and first-person shooter task studies. Toward this end, we re-ran two classic studies (Payne, 2001, and Correll et al., 2002), using modified designs—adding confidence ratings—and analyses, allowing us to determine the shape of the ROCs.[2] Otherwise, we aimed to follow the methods of previous studies, e.g., using the original stimuli (when available) or very similar stimuli. To be clear, we do not claim that in terms of methods, our experiments are direct (exact) replications of previous experiments, in light of some methodological changes. Likewise, our aim was not to see whether the previous results themselves can be replicated. After all, we ran our studies approximately 15 years later using different subject populations, and the nature of racial bias can surely vary. We also know

from the meta-analysis of Mekawi and Bresin (2015) that the prior literature produced varied results. However, in the absence of ROC curves for past studies, we make the working assumption that the *shape* of the ROCs in the present work can shed light on the *shape* of ROCs more generally for weapon identification and shooter task studies. The form of these empirical ROCs will determine the best data analyses for these tasks. Our general prediction is that the ROCs will indeed be curved, putting at risk conclusions from weapon identification and shooter task studies employing ANOVAs on raw scores, PDP, or Quad model analyses.

Experiments 1a and 1b: Priming

Experiments 1a and 1b were modeled on Payne (2001, Experiment 1), which to our knowledge was the original weapon identification task study (as of July 2018, cited nearly 1000 times according to Google Scholar). The general procedure was that participants observed two pictures in quick succession: a face prime (Black or White) then an object (a gun or a tool). Participants responded with a keypress to identify the object as a gun or a tool. We adapted this experimental method to run it in an online environment, and we also collected confidence ratings on a 1–3 scale after each response, so that ROC curves could be plotted. In Experiment 1a, the object was displayed for 200 ms, as in the original study. To avoid ceiling effects from the observed high level of accuracy in Experiment 1a, in Experiment 1b the object was displayed for only 50 ms. Finally, we increased the number of distinct images used as primes and targets from 4 primes (2 Black, 2 White) and 8 targets (4 guns, 4 tools) to 32 (half Black, half White) and 16 (half guns, half tools); our goal was to reduce the possibility that something about the specific prime faces contributed to the observed effects.

Payne's study reported two kinds of analyses on the response rates: an ANOVA on the error responses and PDP modeling. (However, the results of primary interest for Payne's Experiment 1 were in terms of response time. Experiment 2 of that paper introduced a response deadline, leading to a greater focus on the response rates themselves.) As noted, both kinds of analyses on response rates assume linear ROCs. Hence, we also examined the shape of the ROCs to determine whether they were linear or curved.

Method

All experiments reported here were approved by the Institutional Review Board of either the University of Massachusetts (Exps. 1a and 1b) or the University of California Merced (Exps. 2a and 2b).

Participants. The participants in Experiments 1a and 1b were 47 and 75 workers from Amazon's Mechanical Turk (MTurk), respectively, who were at least 18 years of age. These sample sizes were selected to be at least 1.5 times larger than those in Payne's (2001) experiments (his Exp. 1 N = 31; his Exp. 2 N = 32), in anticipation of some data loss from the online format. Based on previous experience, we expected up to 30% of our participants would fail to complete the task as instructed. Even with that level of data loss, our final sample was expected to be at least as large as Payne's. Sample size was determined before any data analysis; as described below, we experienced much lower than expected data loss, yielding relatively large samples (N = 46 and 66). Our focus was on the shape of the ROCs, rather than an expected effect size, making traditional power-based sample size planning less useful. The consistency of the form of these ROCs across our experiments, and across the related literatures described in the General Discussion, suggests that their curvature is not a result of the sample size.

MTurk workers tend to be more demographically diverse than in a typical college student sample, but they provide reliable data that is consistent with lab samples (e.g., Buhrmester, Kwang, & Gosling, 2011). In Experiment 1a, 20 of the 47 participants were female, 26 male, and 1 did not report gender; 35 participants were white, 3 were Asian, 4 African American, 1 Hispanic, 1 Native American, 1 Mixed race, and 1 declined to report. In Experiment 1b, 35 of the 75 participants were male; 51 participants were white, 9 were Asian, 7 African American, 3 Hispanic, and 5 were mixed race.

Stimuli. The stimuli were designed to be similar to those used by Payne (2001). There were 16 White male faces and 16 Black male faces. All of the face prime images were cropped to show only the interior features of the face, and were displayed in black and white with dimensions 400 × 400 pixels (4.17 in^2 on a 96 dpi screen). There were 8 images of guns and 8 images of tools. These target images were shown in black and white with a speckled pattern overlaid, but were still quite clear. The target images were displayed with the same dimensions as the face primes. One hundred forty-four random squares were selected from the target images and randomly mixed to create a backward mask. Face primes and target images were randomly combined to form experimental trials.

Procedure. There were 5 randomly-generated practice trials followed by 192 experimental trials that were evenly distributed across each combination of prime race (Black, White) and target identity (gun, tool). Each trial began with a fixation cross for a random duration between 300 and 700 ms, followed by a face prime image for 200 ms, a target image for either 200 ms (Exp. 1a) or 50 ms (Exp. 1b), and a backward mask that remained visible until the participant made a gun/tool decision using the Q and P keys. Each gun/tool decision was followed by a prompt for a confidence rating on a 3-point scale (1 = Guessing, 2 = Somewhat Confident, 3 = Sure). No other dependent measures were recorded: All measures, manipulations, and exclusions are disclosed for each experiment. No feedback was given; the next trial began 1000 ms after the confidence rating. The sequence of experimental trials was randomly determined for each participant. The experiment was controlled with Qualtrics using the QRTEngine[3] (Exp. 1a; Barnhoom, Hasasnoot, Bocanegra, & van Steenbergen, 2015) or PsiTurk (Exp. 1b; Gureckis et al., 2016) to control display timing. Payne (2001, Exp. 1) did not impose a response deadline, and neither did we, though MTurk workers tend to respond quickly (Smith, Roster, Golden, & Albaum, 2016).

Results

The data of one participant whose ability to discriminate guns from tools was near chance ($d' < 0.5$) and far below the others' was excluded from analysis in Experiment 1a. In Experiment 1b, all data were excluded from 8 participants who had overall d' near chance (<0.5) and for one participant who made the same response ("gun") on all but one trial. We also excluded the first and last trial for every participant in Experiment 1a because of a data recording problem. In addition, individual trials for which the measured target display time was at least 20 ms shorter or longer than the intended display time were excluded from analysis. As a consequence, in Experiment 1a, an average of 3.1% of trials were excluded for each subject; these trials were randomly distributed across conditions. In Experiment 1b, an average of 3.7% of trials were excluded for each subject, randomly distributed across conditions.

In the remaining sections, we first report analyses that assume a linear ROC, then report analyses assuming a curved ROC. We report these analyses for interest, keeping in mind that it is not crucial whether all of the previous findings are replicated. What is crucial for present purposes is the shape of the ROCs themselves.

Analyses that assume a linear ROC

"Gun" response rates. We began by considering the gun/tool responses to the target images. The probability that a "gun" response was elicited by a gun or tool is shown in **Table 2** as a function of both the identity of the target item and the race of the prime face. Participants made more gun responses to guns than to tools in both versions of the experiment (Exp. 1a: $F(1,45) = 53243$, $p < 0.0001$, MSe = .0000, $\eta^2_g = 0.998$; Exp. 1b: $F(1,65) = 705.8$, $p < 0.001$, MSe = .05, $\eta^2_g = 0.825$). However, the race of the prime face did not matter (Exp. 1a: $F(1,45) = 0.287$, $p > 0.59$, MSe = .0003, $\eta^2_g = 0.001$; Exp. 1b: $F(1,65) = 2.085$, $p > .15$, MSe = 0.002, $\eta^2_g = 0.001$) nor was there an interaction of prime race with target object in either experiment (Exp. 1a: $F(1,45) = 0.022$, $p > 0.88$, MSe = .0001, $\eta^2_g = 0.000$; Exp. 1b: $F(1,65) = 0.763$, $p = .386$, MSe = .003, $\eta^2_g = 0.000$).[4]

The ANOVA results for Experiment 1a replicate Payne's: Both show high accuracy and no significant interaction of prime race and target item. Our Experiment 1b differed from Payne's (2001, Exp. 1) in that we reduced decision accuracy by decreasing the display time for the target item whereas Payne (2001, Exp. 2) did so by imposing a response deadline. In his Experiment 2, Payne found that the probability of an erroneous "gun" response was specifically increased by the presentation of a Black prime face, an effect that we did not observe in either version of this priming experiment.[5] One possible explanation of this discrepancy is that our participants were overall less likely to say "gun" than Payne's participants (e.g., $c \approx 0.3$ v. ≤ 0). Because the function relating response bias (c) to "gun" response rates is non-linear, larger bias shifts are required to observe the same magnitude difference in false alarm rates when participants make relatively fewer "gun" responses. Of course, response bias is influenced by factors both under the control of the experimenter and the participants (e.g., Rotello & Macmillan, 2008).

Process dissociation parameter estimates. Following Payne (2001), we also calculated estimates of automatic and controlled processing for each participant according to Jacoby's (1991) process dissociation approach. (This approach has been criticized for other reasons, as we will review briefly in the discussion.) Controlled processing, C, is simply estimated as the difference between correct and incorrect "gun" response rates (C = hit rate – false alarm rate = $H - F$), and automatic processing, A, is estimated as $F/(1-C)$.[6] There were a large number of participants in Experiment 1a who had controlled processing estimates of 1 for at least one of the two prime types, leaving undefined estimates of automatic processing. Rather than correct these values (e.g., Payne, 2001), we evaluated only the Experiment 1b results; the resulting parameter estimates are shown in **Table 3**.

Table 3: Average process dissociation parameter estimates (and standard deviations) from both experiments.

Exp.	Condition	Prime or Agent	Controlled (sd)	Automatic (sd)
1b	50 ms	Black	.72 (.23)	.34 (.28)
		White	.71 (.24)	.22 (.27)
2a	250 ms	Black	.71 (.15)	.37 (.27)
		White	.69 (.14)	.46 (.25)
2b	200 ms	Black	.58 (.15)	.30 (.22)
		White	.56 (.17)	.32 (.19)

Table 2: Average "gun" response rate (and corresponding standard deviation), as well as d' and c (both averaged over subjects), in each condition of Experiments 1 and 2.

Expt	Cond	Black Agent or Prime				White Agent or Prime			
		Gun	Tool	d'	c	Gun	Tool	d'	c
1a	200 ms	0.98 (.03)	0.01 (.03)	*	*	0.99 (.02)	0.01 (.02)	*	*
1b	50 ms	0.82 (.22)	0.08 (.10)	2.68 (.89)	0.27 (.55)	0.79 (.22)	0.08 (.09)	2.59 (.97)	0.30 (.55)
2a	250 ms	0.82 (.11)	0.11 (.09)	2.34 (.73)	0.17 (.33)	0.83 (.10)	0.14 (.10)	2.19 (.61)	0.06 (.33)
2b	200 ms	0.71 (.15)	0.12 (.10)	1.89 (.62)	0.35 (.39)	0.71 (.12)	0.15 (.12)	1.74 (.60)	0.30 (.34)

*Note: Many participants displayed ceiling-level performance, making calculation of d' and c highly dependent on the specific correction factor used for response rates of 0 and 1 for tools and guns, respectively. Thus, we do not report these means.

The estimated proportions of controlled processing exceeded those of automatic processing ($F(1,65) = 117.7$, $p < 0.001$, MSe = 0.083, $\eta^2_g = 0.370$), but there was no effect of the race of the prime face ($F(1,65) = 0.32$, $p > 0.5$, MSe = 0.013, $\eta^2_g = 0.000$) nor an interaction ($F(1,65) = 0.034$, $p > 0.85$, MSe = 0.014, $\eta^2_g = 0.000$). The numerical pattern of results is similar to those reported by Payne (2001, Exp. 1), with essentially identical controlled processing across prime race, and a larger estimate of automatic processing with the Black primes than the White primes, although our effects failed to reach significance.[7]

Analyses that assume a curved ROC

Several studies involving the weapon identification task have reported the signal detection measures of decision accuracy (d' = zHR – zFAR) and response bias (c = –0.5*(zHR + zFAR), see **Table 1**). d' assumes a symmetric and curved ROC, and measures the theoretical distance between the mean strength of evidence provided on gun trials and on tool trials; larger values of d' imply easier discrimination between these two trial types. Larger positive values of the response bias measure c indicate a stronger preference to respond "tool," and larger negative values indicate a stronger bias to say "gun." Key questions in the weapon identification task are whether d' or c depends on the race of the prime. In Experiment 1b, d' is 2.7 when the prime is a Black face, and 2.6 when the prime is White, a non-significant difference: $t(65) = 1.12$, $p > 0.25$ (see **Table 2**), 95% CI for difference = (–0.067, 0.234), $d_z = 0.138$. The response bias measure c also does not differ with the race of the prime in Experiment 1b: 0.27 v. 0.30, $t(65) = 0.36$, $p > 0.7$, 95% CI for difference = (–0.152, 0.106), $d_z = 0.044$. The positive values of c reflect the overall preference of these participants to say "tool" rather than "gun."

Assessments of the ROCs

As Rotello et al. (2015) noted, the form of the ROC for these shooter tasks is unknown, yet the form of the ROC crucially determines which types of analyses are justified. In Experiment 1b, both types of analyses — those that assume linear ROCs (ANOVAs, PDP measures) and those that assume curved and symmetric ROCs (d', c) — resulted in the same conclusion of no significant effect of the race of the prime face on the responses to the gun and tool objects. The observed agreement in these measures is far from assured, as numerous examples in the literature demonstrate (e.g., Snodgrass & Corwin, 1988; Kinchla, 1994; Dube, Rotello, & Heit, 2010; Heit & Rotello, 2014). Moreover, at most one of these classes of analysis can be justified, as the underlying ROCs cannot be both curved and linear simultaneously. Erroneous conclusions are quite likely when inappropriate outcome measures are used. Specifically, the estimated accuracy level is confounded with response bias whenever an inappropriate measure of decision accuracy is applied (e.g., Swets, 1986a; Rotello et al., 2008; Dube et al., 2010; Rotello et al., 2015). For this reason, we turn to our main purpose: the assessment of the ROCs generated by participants in this weapon identification task.

We generated ROCs separately for the trials with Black primes and with White primes. For consistency with the traditional analyses, we defined the hit rate in terms of the probability of a "gun" response to gun images, and the false alarm rate to be the probability of a "gun" response to a tool. The confidence ratings were converted to a 6-point scale ranging from "sure gun" on one end to "sure tool" on the other, allowing us to plot a 5-point ROC. The resulting data are shown in **Figure 1**: the left-most point on the ROC represents "sure gun" response rates, and each point to the right cumulates responses from one additional confidence level (i.e., "sure gun" + "somewhat confident gun", etc.; see Macmillan & Creelman, 2005, for details). Importantly, these ROCs are smoothly curved and appear consistent with a signal detection model.

As can be seen in **Figure 1**, accuracy was very high in Experiment 1a (the ROC is near the top left of the figure) and more moderate in Experiment 1b. Most importantly, the shape of the ROCs is curved rather than linear in both experiments, suggesting that ANOVAs on raw scores, as well as PDP, are inappropriate analyses. Following what is so clear from visual inspection of **Figure 1**, we next used inferential tests to assess curvature at the individual participant level. We assessed the curvature of ROCs in Experiment 1b using their normal-normal twin, the *z*ROC. A *z*ROC can be created by taking the *z*-score of each hit and false alarm rate in the ROC.[8] Conveniently, if the signal detection model's assumptions of curvature in the ROC are correct, then the corresponding *z*ROC is a line, and if the ROC is a line, then the corresponding *z*ROC is curved (Macmillan & Creelman, 2005). We regressed the *z*-scores for the hit rates on the *z*-scores for the false alarm rates, including a quadratic term to assess curvature.[9] For the Black primes, 3 of the 66 participants' individual zROCs (i.e., 4.5%) had significant curvature, as one would expect by chance. For the White primes, none of the 66 *z*ROCs had significant curvature. Overall, then, we conclude from these linear *z*ROCs that the corresponding ROCs are curved; as a consequence, ANOVAs on raw scores and PDP or Quad model analyses of these data are inappropriate.

We also fit the ROCs using maximum likelihood estimation to estimate the model's parameters under several different assumptions about how race might affect the data: 1) it might affect only response bias, which are criterion locations in the model for the confidence levels ($c1_B$–$c5_B$; $c1_W$–$c5_W$); 2) it might affect only the evidence distributions, assessed via their mean locations (d_B, d_W) and variability ($slope_B$, $slope_W$); or 3) it might affect both or 4) neither. Overall, the best-fitting model in both experiments assumed that there were no differences in either accuracy or response bias as a function of prime race; the best-fitting parameters of this model are shown in **Table 4**.[10] Notice that the slope parameters are close to 1 in both experiments, suggesting that the ROCs are approximately symmetric, a finding that justifies the use of d' as a measure of gun/tool discrimination accuracy.

Discussion

The results of Experiments 1a and 1b make several points. First, and most importantly, ROCs collected in the face priming version of the weapon identification task are curved and consistent with the signal detection model.

Table 4: Parameter estimates from the best-fitting signal detection model in each experiment.

Experiment	Condition	Parameters						
		d	*slope*	*c1*	*c2*	*c3*	*c4*	*c5*
1a	200 ms	4.62	1.16	2.84	2.58	2.15	1.77	1.47
1b	50 ms	2.50	0.91	1.88	1.72	1.51	0.89	0.64
2a	Black Agent	2.50	0.69	2.63	1.52	1.21	0.53	–0.36
	White Agent	1.95	1.10	2.04	1.30	1.08	0.56	–0.25
2b	Black Agent	1.88	1.22	2.43	2.56	1.19	0.22	–0.68
	White Agent	1.60	1.03	2.05	1.36	1.04	0.21	–0.58

Note: *d* = mean of target distribution; *slope* = 1/standard deviation of target distribution; *c1*–*c5* = criterion locations for most conservative operating point (*c1*) through most liberal point (*c5*).

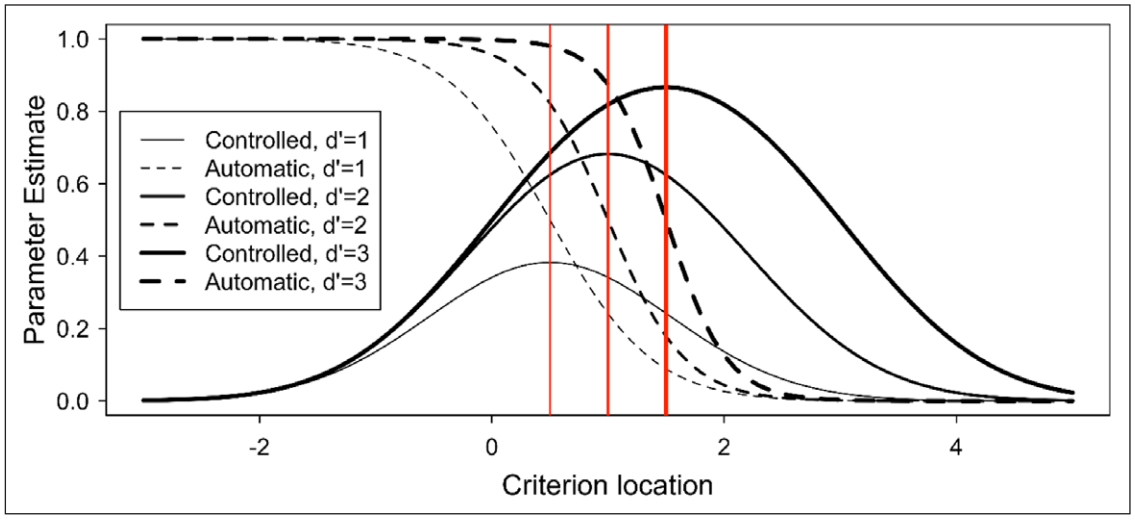

Figure 2: Simulated PDP estimates of controlled (solid curves) and automatic processing (dashed curves) as response bias is varied from conservative to liberal criterion locations (shown right to left on the x-axis), assuming non-gun trials are drawn from a N(0,1) distribution and gun trials are sampled from a N(*d'*,1) distribution. The red vertical lines indicate the neutral criterion location (where SDT's *c* parameter is 0) for three different levels of *d'*, with heavier lines reflecting larger *d'*.

These data are fundamentally inconsistent with the linear ROCs assumed by the ANOVAs on raw scores, and PDP or Quad model analyses, that are commonly reported for weapon identification tasks. In the absence of ROCs from previous studies, we take our results as prima facie evidence that ROCs in the weapon identification task are generally curved. The implications of this point are that previous conclusions based on ANOVAs, Sherman et al.'s (2008) Quad model, and PDP analyses are at risk for having misinterpreted the data because their measures of accuracy are confounded with response bias. While our overall conclusion that there was no effect of race on responses to the target object did not depend on the nature of the analysis for these data, this consistency does not indicate that the ANOVAs or PDP analyses are justified, nor that their conclusions are generally accurate.

For example, consider the PDP analysis of Experiment 1b, which numerically (but not statistically) replicated Payne (2001) in finding larger estimates of automatic processing for Black primes than for White primes. This type of result has been interpreted as a type of implicit bias against Black individuals (Payne, 2001). However, these estimates of automatic processing (A, and those for controlled processing, C) are known to be confounded with response bias (Buchner et al., 1995). Estimates of C vary systematically with response bias, as shown with the solid curves in **Figure 2** for three different possible levels of *d'* (discrimination of gun trials from tool trials). Because estimates of A depend on C, A also varies systematically with response bias, increasing as the criterion becomes more liberal (see the dashed curves in **Figure 2**). In other words, estimates of controlled processing always exceed those for automatic processing for conservative response biases, and PDP estimates of automatic processing are always larger for more liberally-placed decision criteria. If participants respond "gun" more often overall for Black than for White primes, then the estimate of automatic processing will be larger for Black primes. This observation may explain why the literature has typically reported increased estimates of automatic processing for Black agents. Indeed, Payne (2001) reported significantly higher automatic processing with Black primes, and from

his **Table 2** we calculated that his participants were more liberal when responding with Black than White primes (his Exp. 1: c = –0.07 v. 0; Exp. 2: c = –0.17 v. –0.06). In cases where participants say "gun" more often to White agents (as we will see in Experiment 2), the estimate of automatic processing must be numerically greater for White agents. This relationship between response bias and processing estimates is caused by the discrepancy between the assumptions of the analysis (linear ROCs) and the properties of the data (curved ROCs), and has nothing at all to do with the nature of the underlying cognitive process or processes.

Overall, the data of Experiments 1a and 1b reveal no significant effect of the race of the prime on the judgments of the identity of the target object as a gun or a tool. All methods of analysis reached this same conclusion. Despite this unanimity, the ROCs reported in **Figure 1** clearly support the use of signal detection based approaches to analyzing the data (i.e., d' and c). These same ROCs are strongly inconsistent with analyses of shooter bias data that assume linear ROCs, including ANOVAs on error rates, ANOVAs on "gun" response rates, process dissociation methods, and Quad model analyses.

In Experiments 2a and 2b, we generalize these findings by using a different, also classic, paradigm in which an image of a Black or White man holding either a gun or another object is superimposed on a background scene (Correll et al., 2002) and participants are asked to make a rapid response. Analogous to our confidence-rating extensions of Payne's (2001) priming paradigm which allow for examination of the ROCs in Experiments 1a and 1b, in Experiments 2a and 2b, we collected confidence ratings in addition to gun/non-gun responses.

Experiment 2a and 2b: Agents in Context

Experiments 2a and 2b were modeled after Correll et al. (2002, Experiment 1), another early and influential first-person shooter task study (as of July 2018, cited more than 1100 times according to Google Scholar). In this task, participants were shown a picture of a scene such as an interior office, a train station, or a park. After a brief but variable amount of time, an image of a Black or White male actor holding either a gun or some other object, such as a cell phone or soda can, was superimposed on the scene. Participants were instructed to judge quickly whether or not the actor was holding a gun, responding with a keypress whether or not to "shoot." We adapted this experimental method to run it in an online environment, soliciting responses in terms of "gun" or "non-gun" as in Payne (2001), and we also collected confidence ratings on a 1–3 scale after each response, so that ROC curves could be generated. In Correll et al.'s original experiment, responses were allowed for 850 ms after the actor's image was superimposed on the scene. Rather than imposing a time limit, in the present study we limited display time to avoid ceiling effects from a high level of accuracy. Mekawi and Bresin's (2015) meta-analysis, which included studies that did not impose a response deadline, found that the response window did not influence false alarm rates, decision accuracy, or response bias. Moreover, eliminating the response deadline avoids the complication that confidence ratings that follow binary decisions made under time pressure may be based on different information than the binary responses themselves (Petrusic & Baranski, 2003, Van Zandt & Moldanado-Molina, 2004). In Experiment 2a, the actor was displayed for 250 ms. In Experiment 2b, he was displayed for 200 ms. Correll et al.'s study reported two kinds of analyses on the response rates: ANOVAs on the error rates and SDT modeling. As noted previously, these two kinds of analyses make different, incompatible, assumptions about the form of the ROC (linear or curved, respectively). Hence, we also examined the shape of the ROCs to determine whether they were linear or curved.

Method

In Experiments 1a and 1b, a face was used as a prime which could affect subsequent processing of the target item as a gun or tool. In Experiments 2a and 2b, the individual and the item were shown together in a more natural scene. This design looks at real-time paired processing of race and weapons. In addition to making the race and target object stimuli available simultaneously, this paradigm also serves to reduce the prominence of the race factor: participants were told to attend to and respond to the gun/non-gun status[11] of the objects the actors were holding. Their race was not prominently compared, as had been the case with showing faces that are either black or white. Additional differences from the design of Experiments 1a and 1b aligned Experiments 2a and 2b with the study (Correll et al., 2002) from which the materials were taken. These changes consisted of longer time periods between trials, use of fewer trials overall, and displaying guns vs. a range of other objects rather than exclusively tools. The response keys were A and S (instead of Q and P), thus freeing the participant's right hand to make a confidence rating using a mouse-click. Finally, we increased the number of trials to 100 from the 80 used by Correll et al. for increased power.

Participants. The participants in Experiments 2a and 2b were 124 workers from Amazon's Mechanical Turk (MTurk), 62 in each experiment, who were at least 18 years of age. These sample sizes were selected to be about 1.5 times larger than those in the Correll et al. (2002) study (i.e., N = 40), so that our final sample would be at least at large as theirs even with as much as a 30% data loss due to collecting data in an online environment. Sample size was determined before any data analysis, and as is described below, we experienced very little data loss, resulting in relatively large samples (N = 60 and 62). In Experiment 2a, 33 of the participants were male and 29 female. There were 2 Asian, 4 Black, 4 Hispanic, 49 White, and 3 mixed race participants. In Experiment 2b, 26 participants were male and 36 female; 4 Asian, 5 Black, 1 Hispanic, 48 White, and 4 mixed race.

Stimuli. The stimuli were the same as were used by Correll et al. (2002) and were obtained via download at http://psych.colorado.edu/~jclab/FPST.html. There were 100 agent/item/scene pairs; a unique pairing per trial. There were 4 gun images (2 black guns and 2 silver guns) and 4 non-gun images (black wallet, black cellphone, silver cellphone, silver aluminum can). These were held by 10

Black and 10 White male actors. The images of individuals holding an object were superimposed on a set of 20 scenes. 15 scenes were used 4 times, with one full set of race/item type combinations per scene, and 5 were used 8 times with two full sets of combinations. "Blank" versions of the scenes, containing no actors, were displayed both before and after the image of the individual was superimposed; all images were displayed at 640 × 480 pixels.

Procedure. There were 16 practice trials selected from a subset of 20 trial images. The subset contained one example of each background scene each randomly assigned to the four combinations of actor race (Black, White) and object type (gun, non-gun). This was followed by 100 experimental trials evenly distributed across these same four cells, as described previously. Each trial began with a fixation cross for 500 ms, followed by a blank scene[12] for a random duration between 1000 and 4000 ms, followed by the actor image for 250 ms (Exp. 2a) or 200 ms (Exp. 2b), followed by the blank scene again until the participant made a gun/non-gun decision using the A and S keys. Each gun/non-gun decision was followed by a prompt for a confidence rating on a 3-point scale (labeled from left to right on the screen: Very Sure, Sure, Guess) using a mouse click. Reaction times (RTs) were also recorded but not analyzed. Because we had no plans to analyze RTs, we made no effort to insure the accuracy of these RTs in the MTurk environment. No feedback was given; the next trial began 500 ms after a confidence rating was made. The sequence of experimental trials was randomly determined for each participant. The experiments were conducted with Qualtrics using QRTEngine to control display timing.

Results

The data of one participant whose ability to discriminate guns from tools was near chance ($d' < 0.5$) and one participant who made the same confidence rating on every trial were excluded from analysis in Experiment 2a; both were White females. We also excluded individual trials for which the measured target display time was at least 17 ms shorter or longer than the intended display time (i.e., approximately 1 screen refresh cycle). As a consequence, in Experiments 2a and 2b, an average of 1% and 1.3% of trials were excluded, respectively; these trials were randomly distributed over trial types.

Again, we report the findings from analyses assuming a linear ROC then report analyses assuming a curved ROC before turning to the crucial issue of the actual shape of the ROCs.

Analyses that assume a linear ROC

"Gun" response rates. The probability that a "gun" response was elicited by a gun or non-gun is shown in **Table 2** as a function of both the true object identity (gun/non-gun) and the race of the agent. Participants made more gun responses to guns than to other objects in both versions of the experiment (Exp. 2a: $F(1,59) = 1696.0$, $p < 0.001$, MSe = 0.017, $\eta^2_g = 0.925$; Exp. 2b: $F(1,61) = 869.9$, $p < 0.001$, MSe = 0.023, $\eta^2_g = 0.846$). In Experiment 2a, participants were more likely to say "gun" when the agent was White than when he was Black ($F(1,59) = 8.73$, $p < 0.005$, MSe = 0.004, $\eta^2_g = 0.013$), but the effect of agent race was not significant in Experiment 2b: $F(1,61) = 1.46$, $p = 0.23$, MSe = 0.007, $\eta^2_g = 0.003$). There was no interaction of target object with agent race in Experiment 2a ($F(1,59) = 1.49$, $p = 0.23$, MSe = 0.005, $\eta^2_g = 0.003$) and only a marginal effect in Experiment 2b ($F(1,61) = 3.16$, $p = 0.081$, MSe = 0.003, $\eta^2_g = 0.003$).

Process dissociation parameter estimates. Although Correll et al. (2002) did not conduct PDP analyses, for consistency with our description of Experiments 1a and 1b, we report them here as well. The estimated proportions of controlled processing exceeded those of automatic processing (Exp. 2a: $F(1,59) = 70.13$, $p < 0.001$, MSe = 0.068, $\eta^2_g = 0.315$; Exp. 2b: $F(1,61) = 71.9$, $p < 0.001$, MSe = 0.059, $\eta^2_g = 0.331$), but the effect of the agent race was marginal in Experiment 2a ($F(1,59) = 2.96$, $p = 0.09$, MSe = 0.020, $\eta^2_g = 0.006$) and non-significant in Experiment 2b ($F(1,61) = 0.07$, $p = 0.79$, MSe = 0.011, $\eta^2_g = 0.000$). The difference between controlled and automatic processing estimates was smaller for White than for Black agents in Experiment 2a ($F(1,59) = 7.83$, $p < 0.01$, MSe = 0.021, $\eta^2_g = 0.015$). In contrast to Payne (2001), this interaction was due to greater automatic processing to White rather than Black agents. Moreover, the interaction of agent race and parameter type was not found in Experiment 2b ($F(1,61) = 1.91$, $p = 0.17$, MSe = 0.017, $\eta^2_g = 0.004$).

Analyses that assume a curved and symmetric ROC

As for Experiment 1b, we calculated both d' and c as a function of the race of the target. Decision accuracy was marginally higher when the agent was Black rather than White in Experiment 2a (d' = 2.3 v. 2.2, 95% CI for difference = (–0.013, 0.308); $t(59) = 1.84$, $p = 0.071$, $d_z = 0.237$), and significantly so in Experiment 2b (d' = 1.9 v. 1.7, 95% CI for difference = (0.032, 0.279); $t(61) = 2.51$, $p = 0.015$, $d_z = 0.312$). In terms of response bias, participants were more conservative (made fewer "gun" responses) when responding to Black than White agents in Experiment 2a (c = 0.17 v. 0.06, 95% CI for difference = (0.034, 0.184); $t(59) = 2.92$ $p < 0.01$, $d_z = 0.377$), but there was no bias effect in Experiment 2b (c = 0.35 v. 0.30, 95% CI for difference = (–0.034, 0.133); $t(61) = 1.19$, $p > 0.2$, d_z = 0.151). Both of these bias measures are positive, meaning that participants in these experiments had an overall preference to say "not gun" despite the equal numbers of gun and non-gun trials. Although we do not report error-rate based analyses here, a necessary consequence of this conservative bias is that more errors would occur to gun trials than to non-gun trials.

Assessment of the ROCs

The crucial ROCs for Experiments 2a and 2b are shown in **Figure 3**. As in Experiments 1a and 1b, these ROCs are smoothly curved and appear consistent with a signal detection model. They are clearly not linear, implying that ANOVAs based on raw scores, as well as PDP analyses, are inappropriate even when they happen to reach the same conclusions as the signal detection analyses, as occurred in Experiments 1a and 1b. We confirmed this visual impression with an analysis of the *z*ROCs, as

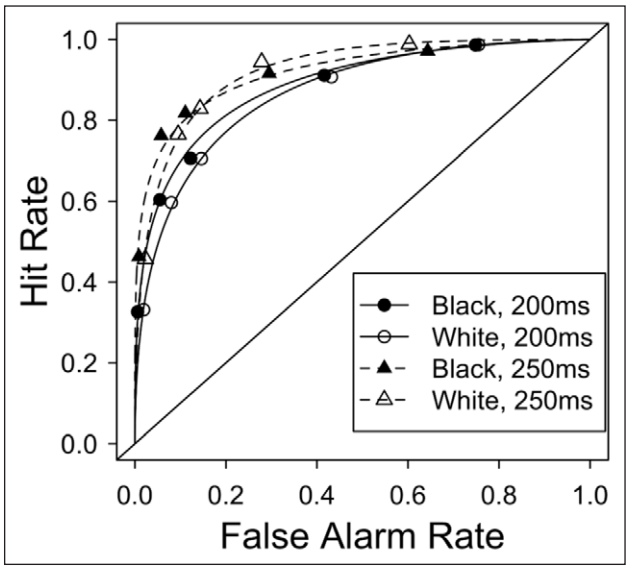

Figure 3: Receiver operating characteristic curves (ROCs) for the data from Experiments 2a (250 ms displays) and 2b (200 ms). The symbols indicate the observed response rates at each level of decision confidence; the curved show the best-fitting function when fit to each ROC independently.

in Experiment 1b. In Experiment 2a, 2 participants had *z*ROCs for the White agents with significant curvature (3.4%; one participant's *z*ROC could not be assessed for curvature because it contained too few points), and 2 had significantly curved *z*ROCs for Black agents (3.5%; two participants' *z*ROCs could not be assessed). In Experiment 2b, one participant had a significantly curved *z*ROC for the White agents (1.7%; 4 could not be assessed) and one participant had a significantly curved *z*ROC for the Black agents (1.7%; 2 could not be assessed). Overall, about 88% of the individual participants' *z*ROCs in Experiments 2a and 2b are linear and consistent with the curved ROCs predicted by the SDT model.

Accuracy (readily observed as the area under each curve) appears higher for the longer display times of Experiment 2a (250 ms) than those of Experiment 2b (200 ms), as would be expected; differences are slight as a function of actor race. As in Experiments 1a and 1b, the visual impression offered by these ROCs is that the race of the actor had relatively little influence on the responses.

To assess these ROCs statistically, we fit them with the same models as in Experiments 1a and 1b. In Experiment 2a, restricting the evidence distributions to have the same mean and standard deviation for both agent races reduced the model fit significantly: $\Delta G^2(2) = 45.62$, $p < 0.001$. However, this difference is attributable to the difference in the symmetry of the ROC only: the accuracy of the gun/non-gun decisions, measured by the area under the curve (e.g., Macmillan & Creelman, 2005), is 0.92 in both cases. These areas do not differ reliably, of course (D = 0.48, $p > 0.6$, analyzed with the pROC package in R, Xavier et al., 2011). Assuming there was no effect of agent race on response bias also impaired the quality of fit ($\Delta G^2(5) = 38.56$, $p < 0.001$); participants tended to respond "gun" more often for White than Black agents. Thus, in Experiment 2a, we conclude that actor race influenced only the response bias of the participants: they were more likely to give a "gun" response when the actor was White than when he was Black, but the accuracy of their responses did not differ as a function of race.

The results are similar for Experiment 2b. Constraining the mean and standard deviation parameters to be the same for both agent races impaired the fit according to the nested model comparison ($\Delta G^2(2) = 11.71$, $p < 0.01$), although the difference in decision accuracy measured with area under the curve (0.89 Black v. 0.87 White) was only marginally significant (D = 1.77, $p = 0.077$). Restricting the response criteria to be equal for both agent races significantly reduced the fit ($\Delta G^2(5) = 21.80$, $p < 0.001$); participants made more "gun" responses when the agent was White than when he was Black. Parameter estimates from the best-fitting (fully-free) model are shown in **Table 4**.

Discussion

As in Experiments 1a and 1b, Experiments 2a and 2b revealed curved ROCs. This finding is consistent with the signal detection analysis of the first-person shooter task in terms of *d'* and *c*, and inconsistent with analyses that assume linear ROCs (e.g., ANOVA, PDP). Again, in the absence of information from prior studies, we take the present finding as prima facie evidence that shooter task ROCs are generally curved. The signal detection interpretation of the present data is that decision accuracy is higher when the agent is Black than when he is White, according to *d'* (marginally so in Experiment 2a; significantly in 2b), but that the area under the ROC was only marginally higher in Experiment 2b. Experiment 2a revealed a significantly greater rate of "gun" responses to White than Black agents, and this response bias difference led to the expected increase in the PDP estimate of automatic processing for the White agents (see **Figure 2**). However, neither of these effects were found in Experiment 2b. Estimating automatic and controlled processing in Correll et al.'s (2008) data from their Table 1, we find that controlled processing exceeded automatic for both White and Black agents (C = 0.91 in both cases). Consistent with the more liberal response criterion for Black than White agents (*c* = –0.29 v. –0.13) in Correll et al.'s experiment, the estimate of automatic processing is also greater for Black than White agents (A = 0.78 v 0.63). These results support our argument that the discrepancy between the assumptions of the PDP model (linear ROCs) and the data (curved ROCs) determines the automatic and controlled processing estimates; those estimates do not depend on differences in cognitive processes.

General Discussion

The major contribution of these experiments is to identify the form of the empirical ROC across two basic tasks investigating shooter bias, namely a weapon identification task and a first-person shooter task. In all cases, the observed ROC is strongly curved and consistent with the assumptions of the signal detection model, indicating that of the previously reported analysis methods, SDT analyses

are appropriate and justified, whereas the other analyses are not justified and confound decision accuracy with response bias. In the absence of any evidence for linear ROCs, and with evidence from four new experiments employing two different tasks revealing curved ROCs, it seems likely that if previous studies of these tasks had also plotted ROCs, they would have also been curved. Thus, previous conclusions based on the signal detection measures *d'* and *c* provide the most reliable and best interpretation of the prior literature (e.g., Swets, 1986a; Rotello et al., 2008; Rotello et al., 2015). Turning back to **Table 1**, it appears that 21 of 45 studies (47%), or about half of all published experiments on shooter bias, did not employ the form of analysis that is justified, because they only reported analyses that assume a linear ROC, analyses that appear to be at risk of leading to faulty conclusions such as misinterpreting a response bias effect as a difference in decision accuracy. Of the 24 remaining studies with SDT analyses, 16 of them additionally reported either ANOVAs on response proportions, and/or PDP or Quad modeling. Hence these 16 studies reported a mixture of justified and at-risk analyses. We have only found 8 published studies without any analyses on response rates that these ROCs now indicate are at-risk. Thus, what had seemed to be an area of social psychology with a broad empirical base for drawing scientific and practical conclusions instead may be at risk of having drawn unfounded inferences overall. Given that our primary aim has been to establish which statistical analyses are appropriate, here we do not attempt to resolve what seem to be conflicting results in the literature, including those that have relied on at-risk analyses.

As noted in the introduction, our concerns were about analyses of response rates rather than response times, which are widely reported for these tasks (in **Table 1**, 32 out of 45, or 71% of studies). Our argument about shape of ROCs does not apply to response time data, so on that basis we make no specific claims about the appropriateness of response time analyses. However, in 26 of these 32 studies, the response time analyses were combined with other analyses on response rates that our new ROC data indicate are inappropriate (ANOVAs and/or PDP/Quad modeling). So, the literature has been informed by a mixture of analyses that our ROCs show are sometimes justified and sometimes problematic.

With these points made, we also urge caution when interpreting the response time analyses, for several reasons unrelated to ROCs. First, the observed effect sizes are small, even when assessed separately for gun (mean d = 0.13, Mekawi & Brelin, 2015) and non-gun trials (mean d = 0.11). Second, only 6 of the 32 reaction time analyses summarized by Mekawi and Brelin (2015) revealed significant race effects for *both* gun and non-gun trials. These reaction time effects, even if taken at face value, have more than one interpretation in terms of psychological processing; additional empirical and modeling work is needed to pinpoint the cause. A good start in this direction was offered by Correll, Wittenbrink, Crawford, and Sadler (2015), who analyzed reaction time data from 3 shooter tasks using a simplified diffusion model. They found that information processing was speeded for stereotype-consistent trials (black/gun, white/non-gun) compared to inconsistent trials. However, the number of trials was relatively low for this type of modeling (van Ravenzwaaij & Overauer, 2009; Lerche, Voss, & Nagler, 2017). More recently, Pleskac, Cesario, and Johnson (*in press*) showed that the rate of information accrual depends on the race of the target as well as the nature of the object he holds, potentially speeding decisions for Black targets. However, their participants sometimes required more information to make a "shoot" decision about Black targets than White targets, effectively offsetting the processing speed difference. This type of modeling is crucial for identifying underlying causes of any RT differences.

Overall, then, we are reluctant to draw larger conclusions based on the body of all published studies of the weapon identification and first-person shooter tasks. However, here we consider the implications of our own experiments. In four experiments that extended two classic paradigms (Payne, 2001; Correll et al., 2002), we observed little to no evidence for race-based effects on gun/non-gun decision accuracy. This was true both when decision accuracy was very high (Experiments 1a) and when it was lower (Experiments 1b, 2a, and 2b). We also observed no evidence that supports a negative bias toward Black agents or primes: the "gun" response rate was never significantly higher for stimuli involving Black individuals than those involving White individuals, and in Experiment 2a the "gun" response rate to Black agents was significantly lower. Thus, these data are consistent with the conclusions of Mekawi and Bresin's (2015) meta-analysis on the effects of race on decision accuracy, but the data contradict their conclusion that participants are generally more likely to "shoot" at Black agents. An optimistic view of these different empirical outcomes is that societal changes in the 17 years since publication of Payne's initial study have weakened the association between guns and Black agents. A more realistic view might be simply that response biases vary readily (e.g., Rotello & Macmillan, 2008). Indeed, Mekawi and Bresin's (2015) meta-analysis included 13 (of 29) individual experiments in which no difference in response bias was observed across agent race (like our Exps. 1a, 1b, and 2b), and another 3 experiments which (like our Exp. 2a) reported a significant bias to "shoot" more at White agents.

More critically, these experiments provide important information about the most appropriate analyses of data from first-person shooter and weapon identification tasks. Traditional treatments have evaluated either the "gun" response rate or error response rate to Black and White agents holding guns or non-guns with an ANOVA. These analyses are based on difference scores that assume that the relationship between the hit rate and the false alarm rate (i.e., the ROC) is linear. For example, the interaction effect of race with the target object is the difference of differences in "gun" response rates to guns and tools: $(H - F)_{Black} - (H - F)_{White}$, a statistic that depends critically on the assumption that H – F is constant for all response biases. When that assumption is violated, as it is when the empirical ROCs are curved, then significant interaction

effects can be observed even in the absence of any decision accuracy differences across conditions (e.g., Dube et al., 2010; Heit & Rotello, 2014).

Another common analysis of data from these tasks is to estimate the contributions of putative controlled and automatic processes to the decisions, using process dissociation (PDP: Jacoby, 1991). Two findings are typical: 1) estimates of controlled processing exceed estimates for automatic processing for both Black and White agents, and 2) estimates of automatic processing are higher for Black than for White agents. As **Figure 2** demonstrates, both of these effects can be attributed to the confounding of the parameter estimates with response bias differences across agent race. Because the PDP technique relies on an assumption of a linear ROC (Buchner et al., 1995; Pazzaglia et al., 2013), the resulting process estimates vary systematically with response bias whenever the empirical ROC is curved and consistent with a signal detection model. **Figure 2** shows that, under these conditions, conservative response biases (i.e., relatively low "gun" response rates) necessarily yield higher PDP estimates of controlled than automatic processing, and that more liberal response biases necessarily yield higher estimates of automatic processing (see also Ratcliff, Van Zandt, & McKoon, 1995). Thus, researchers should be careful to interpret PDP estimates in the literature as reflective of response bias effects, not cognitive processes.

Potential limitations

In light of our strongly supported conclusion that ROCs for the weapon identification task and the shooter task are curved based on the current experiments, we pause to consider possible concerns about this conclusion. Because previous studies did not collect or report confidence ratings, we can only infer the likely shape of ROCs that would have been obtained. The previous studies used different visual stimuli (other than our use of pictures from Correll et al. 2002 in Exps. 2a and 2b). However, a great deal of previous research using signal detection measures suggests that empirical shape of ROCs is robust over choice of materials. For example, in numerous studies of recognition memory as well as visual perception, ROCs have been found to be curved using pictures of faces (e.g., Dobolyi & Dodson, 2013), scenes (e.g., Evans, Rotello, Li, & Rayner, 2009), schematic line drawings of objects (Dube & Rotello, 2012), as well as word lists (e.g., Dube & Rotello, 2012; see Rotello, 2017, for a review of recognition memory ROCs). Moreover, the empirical shape of ROCs is robust when looking across experiments run online versus in the lab. For example, most studies of recognition memory have been lab-based, but eyewitness recognition experiments have consistently observed curved ROCs whether run online (e.g., Mickes et al., 2017) or in the lab (e.g., Mickes, Flowe, & Wixted, 2012). Finally, as we adapted our studies for an online environment, there were minor changes in timing (time of stimulus presentation and interval allowed for response). Yet the shape of ROCs appears to be consistent across such changes in methodology as well, including instructions that emphasize speed versus accuracy (e.g., Osth, Bora, Dennis, & Heathcote, 2017).

We next turn to a related issue, whether the very fact that we ask for confidence ratings unduly changes the nature of the task. It is important to state that there is ample evidence that adding confidence ratings does not have a disturbing effect on experimental results, and also that such confidence ratings are very likely to be based on the same psychological evidence as the binary decision for participants' binary responses (see, e.g., Petrusic & Baranski, 2003, Van Zandt & Moldanado-Molina, 2004, for decisions made without a deadline). For example, Egan, Shulman, & Greenberg (1959) ran an auditory perception experiment both with and without confidence ratings, finding essentially identical results. In a recognition memory task, Benjamin, Tullis, & Lee (2013) found that the number of available confidence levels had no obvious effect on the shape of the ROC. Likewise, in an eyewitness identification study, Mickes et al. (2017) found that recognition ROCs were similar regardless of whether they were generated from confidence ratings or from an old-new decision with instructional manipulations of response bias. Note that asking for confidence ratings, e.g., on a 1–3 scale, is different than asking for justifications or reasons, e.g., as in Nisbett and Wilson (1977). Again, it is also important to emphasize that our goal here was not to determine whether all of the results from previous weapon identification and shooter task studies can be replicated, but rather to determine the shape of ROCs in these tasks. Keeping in mind that collecting confidence ratings is intrinsic to our method of testing assumptions of analyses, we acknowledge that, strictly speaking, our conclusions are limited to the experimental conditions we ran. For example, in our experiments with confidence ratings, we concluded that ANOVA was an inappropriate analysis for binary judgments. As we have claimed, this finding puts other studies—ones that have used similar methods other than confidence ratings and that have not tested the assumptions of their analyses—at risk of having not appropriately controlled Type I error rates if they also relied on ANOVA. In the context of **Table 1**, none of these published studies tested the assumptions of the analyses they utilized. Our results call the ANOVAs from past studies into question. Note that many studies reported both ANOVAs and SDT analyses. Because these two types of analyses make contradictory assumptions, it would be logically impossible for both analyses to be correct for the same study.

Whether the overall conclusions in the shooter bias literature are justified is a question that will only be answered with additional analyses of existing studies, and the inclusion of other ROC-based experiments. The signal-detection based summary provided by Mewaki and Bresin (2015) indicates that the most accurate current interpretation of the literature is that race plays no significant role in decision accuracy in these shooter tasks, and that the effects of response bias, if any, are highly variable across both participants and experiments (see their Figure 3) for reasons that are not well-understood. Those bias differences imply differences in PDP- or

Quad-model-based estimates of automatic and controlled processes, as we described previously.

Conclusion

Although our focus here is on the weapon identification and first-person shooter tasks, testing the assumptions behind analyses is of broad relevance to social psychology research. Whenever there is a study involving response proportions, the researcher has the choice of analyzing the raw scores (assuming linear ROCs) or converting them to *z*-scores and applying SDT (assuming curved ROCs). Both kinds of analyses cannot be correct for the same data set. In addition to the weapon identification task, Rotello et al. (2015) discussed three other widely-studied research topics — eyewitness memory, belief bias in reasoning, and referral for suspected child maltreatment — each of which could be affected by social categories or stereotypes, and each of which features a long-standing literature that could include many unjustified analyses. The more general point is that it is always important to scrutinize the measures being used, and to choose statistical summaries of performance that are justified by the properties observed in the data (such as curvature in the ROC). We encourage the use of signal-detection based measures, such as *d'* and *c*, in the analysis of these tasks.

With replication studies being encouraged, particularly running the same or similar studies with the same analyses, there is a risk of cumulatively building science and policy based on a faulty foundation. If unjustified analyses were originally used and subsequently repeated without critical analysis, such studies would potentially increase researchers' confidence in incorrect conclusions. As noted by Brandt et al. (2014), there is the risk that "[i]f the original study was plagued by confounds or bad methods, then the replication study will similarly be plagued by the same limitations" (p. 222). More optimistically, in a research culture where data are shared and are open for re-analysis, and where there is greater attention paid to methodological choices including choice of dependent measures, there is a greater chance of drawing justified conclusions from experimental data, better informing important questions for science and society.

Data Accessibility Statement

All participant data can be found on this paper's project page on the Open Science Framework: https://osf.io/fzt78/?view_only=ceb044d34d414a4fa0817f457a138a5f.

Notes

[1] The risk of a Type II error is less well-studied but appears to be similar across dependent measures of decision accuracy (Rotello et al., 2008).

[2] Although the point that ANOVAs in these tasks assume linear ROCs is made by necessity and is unchanging, in general, models that predict a linear ROC can be modified, post-hoc, to fit a curved ROC when that ROC is generated from confidence ratings (e.g., Malmberg, 2002). Strictly speaking, we are examining the PDP or Quad models as proposed rather than hypothetical modifications that have yet to be developed. A number of published papers address this issue, and we will not repeat all of the arguments here (for a review, see Pazzaglia, Dube, & Rotello, 2013).

[3] In September of 2016, a Qualtrics update led to the discontinuation of QRTEngine. All of our QRTEngine experiments were run in the summer of 2015, prior to that change.

[4] Judd, Westfall, & Kenny (2012) showed that this interaction effect can be overestimated when the stimuli are not treated as having random effects. Their conclusion is not of central concern here, for three reasons. First, none of the interaction effects we observe are significant at $\alpha = .05$. Second, ANOVA is not an appropriate analysis method for response rate data like these, as Dube et al. (2010) demonstrated; we include these analyses only for consistency with the literature. Third, in the context of the ROCs that are our primary interest, there is no averaging over trials: stimulus variability is reflected in the shape and height of the ROC itself (see Macmillan & Creelman, 2005).

[5] In addition, a pilot version of Experiment 1b (N = 22) reached the same conclusions as Experiments 1a and 1b for every analysis.

[6] Note that this *C* parameter is not the same as the lower case *c* parameter in SDT analyses.

[7] Payne (2001) reported separate simple effects tests for controlled and automatic processing, finding a significant effect of prime only on automatic processing. Those tests also failed to reach significance in Experiment 1b (automatic: $t(65) = 0.20$, $p > 0.84$; controlled: $t(65) = 0.85$, $p > 0.39$).

[8] We did not analyze Experiment 1a's data this way because the high accuracy level of most participants resulted in many undefined z-scores. In Experiment 1b, a log-linear correction was applied for 0s and 1s (Snodgrass & Corwin, 1988; Rotello et al., 2008).

[9] Technically, this analysis is problematic because it fails to account for error variability in the false alarm rates. However, Ratcliff, McKoon, and Tindall (1994) found that least-squares fits to ROCs were generally comparable to fits derived with maximum-likelihood estimation techniques.

[10] In Experiment 1a, the quality of the fit was not significantly reduced when accuracy ($\Delta G^2(2) = 0.69$, $p = 0.71$), criterion locations ($\Delta G^2(5) = 1.35$, $p = 0.93$), or both ($\Delta G^2(7) = 1.49$, $p = 0.98$) were held constant across the race of the prime. Likewise, in Experiment 1b, constraining accuracy ($\Delta G^2(2) = 1.18$, $p = 0.55$), response criterion locations ($\Delta G^2(5) = 5.43$, $p = 0.37$), or both ($\Delta G^2(7) = 8.11$, $p = 0.32$) did not reduce the fit statistically.

[11] The original task used "shoot" and "don't shoot" as the response options.

[12] The original Correll et al. procedure used a series of four different blank scenes across the same length of time. Due to a misinterpretation of the original procedure, we used the same scene as the actor would subsequently appear within.

Acknowledgements

This material is based on work done while E.H. was serving at the National Science Foundation. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We thank Colin Quirk for help with Experiments 1a and 1b as well as a pilot study.

Competing Interests

The authors have no competing interests to declare.

Author Contributions

- These experiments were conceived and designed by EH and CMR; Experiment 2 was programmed by LJK, who also contributed to its design.
- CMR and EH wrote the manuscript; LJK provided important revisions.
- CMR and LJK were responsible for data collection and analysis.
- All authors were involved in data interpretation and have approved this manuscript.

References

Akinola, M., & **Mendes, W. B.** (2012). Stress-induced cortisol facilitates threat-related decision making among police officers. *Behavioral Neuroscience, 126*, 167–174. DOI: https://doi.org/10.1037/a0026657

Amodio, D. M., Harmon-Jones, E., Devine, P. G., Curtin, J. J., Hartley, S. L., & **Covert, A. E.** (2004). Neural signals for the detection of unintentional race bias. *Psychological Science, 15*, 88–93. DOI: https://doi.org/10.1111/j.0963-7214.2004.01502003.x

Barnhoom, J. S., Hasasnoot, E., Bocanegra, B. R., & **van Steenbergen, H.** (2015). QRTEngine: An easy solution for running online reaction time experiments using Qualtrics. *Behavior Research Methods, 47*, 918–929. DOI: https://doi.org/10.3758/s13428-014-0530-7

Benjamin, A. S., Tullis, J. G., & **Lee, J. H.** (2013). Criterion noise in ratings-based recognition: Evidence from the dffects of response scale length on recognition accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 5*, 1601–1608. DOI: https://doi.org/10.1037/a0031849

Brandt, M. J., Ijzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J. R., & **Van't Veer, A.** (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50*, 217–224. DOI: https://doi.org/10.1016/j.jesp.2013.10.005

Brewer, G. A., Ball, B. H., & **Ware, J. M.** (2016). Individual differences in working memory capacity and shooting behavior. *Journal of Applied Research in Memory and Cognition, 5*, 185–191. DOI: https://doi.org/10.1016/j.jarmac.2016.04.004

Buchner, A., Erdfelder, E., & **Vaterrodt-Plünnecke, B.** (1995). Toward unbiased measurement of conscious and unconscious memory processes within the process dissociation framework. *Journal of Experimental Psychology: General, 124*, 137–160. DOI: https://doi.org/10.1037/0096-3445.124.2.137

Buhrmester, M., Kwang, T., & **Gosling, S. D.** (2011). Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality Data? *Perspectives on Psychological Science, 6*, 3–5. DOI: https://doi.org/10.1177/1745691610393980

Conrey, F. R., Gawronski, B., Sherman, J. W., Hugenberg, K., & **Groom, C. J.** (2005). Separating multiple processes in implicit social cognition: The quad model of implicit task performance. *Journal of Personality and Social Psychology, 89*, 460–487. DOI: https://doi.org/10.1037/0022-3514.89.4.469

Correll, J. (2008). $1/f$ noise and effort on implicit measures of bias. *Journal of Personality and Social Psychology, 94*, 48–59. DOI: https://doi.org/10.1037/0022-3514.94.1.48

Correll, J., Park, B., Judd, C. M., & **Wittenbrink, B.** (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology, 83*, 1314–1329. DOI: https://doi.org/10.1037//0022-3514.83.6.1314

Correll, J., Park, B., Judd, C. M., & **Wittenbrink, B.** (2007b). The influence of stereotypes on decisions to shoot. *European Journal of Social Psychology, 37*, 1102–1117. DOI: https://doi.org/10.1002/ejsp.450

Correll, J., Park, B., Judd, C. M., Wittenbrink, B., Sadler, M. S., & **Keesee, T.** (2007a). Across the thin blue line: Police officers and racial bias in the decision to shoot. *Journal of Personality and Social Psychology, 92*, 1006–1023. DOI: https://doi.org/10.1037/0022-3514.92.6.1006

Correll, J., Urland, G. R., & **Ito, T. A.** (2005). Event-related potentials and the decision to shoot: The role of threat perception and cognitive control. *Journal of Experimental Social Psychology, 42*, 120–128. DOI: https://doi.org/10.1016/j.jesp.2005.02.006

Correll, J., Wittenbrink, B., Crawford, M. T., & **Sadler, M. S.** (2015). Stereotypic vision: How stereotypes disambiguate visual stimuli. *Journal of Personality and Social Psychology, 108*, 219–233. DOI: https://doi.org/10.1037/pspa0000015

Correll, J., Wittenbrink, B., Park, B., Judd, C. M., & **Goyle, A.** (2011). Dangerous enough: Moderating racial bias with contextual threat cues. *Journal of Experimental Social Psychology, 47*, 184–189. DOI: https://doi.org/10.1016/j.jesp.2010.08.017

Cox, W. T. L., Devine, P. G., Plant, E. A., & **Schwartz, L. L.** (2014). Toward a comprehensive understanding of officers' shooting decisions: No simple answers to this complex problem. *Basic and Applied Social Psychology, 36*, 356–364. DOI: https://doi.org/10.1080/01973533.2014.923312

Dobolyi, D. G., & **Dodson, C. S.** (2013). Eyewitness confidence in simultaneous and sequential lineups: A criterion shift account for sequential mistaken identification overconfidence. *Journal of Experimental Psychology: Applied, 19*(4), 345–357. DOI: https://doi.org/10.1037/a0034596

Dube, C., & **Rotello, C. M.** (2012). Binary ROCs in perception and recognition memory are curved. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 130–151. DOI: https://doi.org/10.1037/a0024957

Dube, C., Rotello, C. M., & **Heit, E.** (2010). Assessing the belief bias effect with ROCs: It's a response bias effect. *Psychological Review, 117*, 831–863. DOI: https://doi.org/10.1037/a0019634

Egan, J. P. (1958). Recognition Memory and the Operating Characteristic (Technical Note AFCRC-TN-58-51). Indiana University Hearing and Communication Laboratory, Bloomington, IN.

Egan, J. P., Schulman, A. I., & **Greenberg, G. Z.** (1959). Operating characteristics determined by binary decisions and by ratings. *Journal of the Acoustical Society of America, 31*, 768–773. DOI: https://doi.org/10.1121/1.1907783

Evans, K., Rotello, C. M., Li, X., & **Rayner, K.** (2009). Scene perception and memory revealed by eye movements and ROC analyses: Does a cultural difference truly exist? *Quarterly Journal of Experimental Psychology, 62*, 276–285. [PMCID: PMC2668147]. DOI: https://doi.org/10.1080/17470210802373720

Fleming, K. K., Bandy, C. L., & **Kimble, M. O.** (2010). Decisions to shoot in a weapon identification task: The influence of cultural stereotypes and perceived threat on false positive errors. *Social Neurocience, 5*, 201–220. DOI: https://doi.org/10.1080/17470910903268931

Govorun, O., & **Payne, B. K.** (2006). Ego—depletion and prejudice: Separating automatic and controlled components. *Social Cognition, 24*, 111–136. DOI: https://doi.org/10.1521/soco.2006.24.2.111

Greenwald, A. G., & **Banaji, M. R.** (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological Review, 102*, 4–27. DOI: https://doi.org/10.1037/0033-295X.102.1.4

Greenwald, A. G., McGhee, D. E., & **Schwartz, J. L.** (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology, 74*, 1464–1480. DOI: https://doi.org/10.1037/0022-3514.74.6.1464

Greenwald, A. G., Oakes, M. A., & **Hoffman, H. G.** (2003). Targets of discrimination: Effects of race on responses to weapons holders. *Journal of Experimental Social Psychology, 39*, 399–405. DOI: https://doi.org/10.1016/S0022-1031(03)00020-9

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., Halpern, D., Hamrick, J. B., & **Chan, P.** (2016). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods, 48*, 829–842. DOI: https://doi.org/10.3758/s13428-015-0642-8

Harmer, S. (2012). The role of multiple racial stereotypes in simulated shooting decisions. Carleton University.

Heit, E., & **Rotello, C. M.** (2014). Traditional difference-score analyses of reasoning are flawed. *Cognition, 131*, 75–91. DOI: https://doi.org/10.1016/j.cognition.2013.12.003

Hunsinger, M. (2011). Threat on the mind: The impact of incidental fear on race bias in rapid decision-making. Order No. AAI3427537, *Dissertation Abstracts International. Section B: The Sciences and Engineering, 594*.

Huntsinger, J. R., Sinclair, S., & **Clore, G. L.** (2009). Affective regulation of implicitly measured stereotypes and attitudes: Automatic and controlled processes. *Journal of Experimental Social Psychology, 45*, 560–566. DOI: https://doi.org/10.1016/j.jesp.2009.01.007

Ito, T. A., Friedman, N. P., Bartholow, B. D., Correll, J., Loersch, C., Altamirano, L. J., & **Miyake, A.** (2015). Toward a comprehensive understanding of executive cognitive function in implicit racial bias. *Journal of Personality and Social Psychology, 108*, 187–218. DOI: https://doi.org/10.1037/a0038557

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30*, 513–541. DOI: https://doi.org/10.1016/0749-596X(91)90025-F

James, L., Klinger, D., & **Vila, B.** (2014). Racial and ethnic bias in decisions to shoot seen through a stronger lens: Experimental results from high-fidelity laboratory settings. *Journal of Experimental Criminology, 10*, 323–340. DOI: https://doi.org/10.1007/s11292-014-9204-9

James, L., Vila, B., & **Daratha, K.** (2013). Results from experimental trials testing participant resources to White, Hispanic and Black suspects in high-fidelity deadly force judgment and decision-making simulations. *Journal of Experimental Criminology, 9*, 189–212. DOI: https://doi.org/10.1007/s11292-012-9163-y

Judd, C. M., Westfall, J., & **Kenny, D. A.** (2012). Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103*, 54–69. DOI: https://doi.org/10.1037/a0028347

Kahn, K. B., & **Davies, P. G.** (2010). Differentially dangerous? Phenotypic racial stereotypicality increases implicit bias among ingroup and outgroup members. *Group Processes & Intergroup Relations, 14*, 569–580. DOI: https://doi.org/10.1177/1368430210374609

Kinchla, R. A. (1994). Comments on Batchelder and Riefer's multinomial model for source monitoring. *Psychological Review, 101*, 166–171. DOI: https://doi.org/10.1037/0033-295X.101.1.166

Klauer, K. C., & **Voss, A.** (2008). Effects of race on responses and response latencies in the weapon identification task: A test of six models. *Personality and Social Psychology Bulletin, 34*, 1124–1140. DOI: https://doi.org/10.1177/0146167208318603

Kleider, H. M., & **Parrott, D. J.** (2009). Aggressive shooting behavior: How working memory and threat influence shoot decisions. *Journal of Research in Personality,*

43, 494–497. DOI: https://doi.org/10.1016/j.jrp.2008.12.007

Kleider, H. M., Parrott, D. J., & **King, T. Z.** (2010). Shooting behavior: How working memory and negative emotionality influence police officer shoot decisions. *Applied Cognitive Psychology, 24*, 707–717. DOI: https://doi.org/10.1002/acp.1580

Lambert, A. J., Payne, B. K., Jacoby, L. L., Shaffer, L. M., Chasteen, A. L., & **Khan, S. R.** (2003). Stereotypes as dominant responses: On the "social facilitation" of prejudice in anticipated public contexts. *Journal of Personality and Social Psychology, 84*, 277–295. DOI: https://doi.org/10.1037/0022-3514.84.2.277

Lerche, V., Voss, A., & **Nagler, M.** (2017). How many trials are required for parameter estimation in diffusion modeling? A comparison of different optimization criteria. *Behavior Research Methods, 49*, 513–537. DOI: https://doi.org/10.3758/s13428-016-0740-2

Ma, D. S., & **Correll, J.** (2011). Target prototypicality moderates racial bias in the decision to shoot. *Journal of Experimental Social Psychology, 47*, 391–396. DOI: https://doi.org/10.1016/j.jesp.2010.11.002

Ma, D. S., Correll, J., Wittenbrink, B., Bar-Anan, Y., Sriram, N., & **Nosek, B. A.** (2013). When fatigue runs deadly: The association between fatigue and racial bias in the decision to shoot. *Basic and Applied Social Psychology, 35*, 515–524. DOI: https://doi.org/10.1080/01973533.2013.840630

Macmillan, N. A., & **Creelman, C. D.** (2005). *Detection theory: A users guide* (2nd ed.). New York: Cambridge University Press.

Malmberg, K. J. (2002). On the form of ROCs constructed from confidence ratings. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 380–387. DOI: https://doi.org/10.1037/0278-7393.28.2.380

Mason, I. (1982). A model for assessment of weather forecasts. *Australian Meteorological Magazine, 30*, 291–303.

Mekawi, Y., & **Bresin, K.** (2015). Is the evidence from racial bias shooting task studies a smoking gun? Results from a meta-analysis. *Journal of Experimental Social Psychology*, *61*, 120–130. DOI: https://doi.org/10.1016/j.jesp.2015.08.002

Mekawi, Y., Bresin, K., & **Hunter, C. D.** (2016). White fear, dehumanization, and low empathy: Lethal combinations for shooting biases. *Cultural diversity and ethnic minority psychology, 22*, 322–332. DOI: https://doi.org/10.1037/cdp0000067

Mickes, L., Flowe, H. D., & **Wixted, J. T.** (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous versus sequential lineups. *Journal of Experimental Psychology: Applied, 18*, 361–376. DOI: https://doi.org/10.1037/a0030609

Mickes, L., Seale-Carlisle, T. M., Wetmore, S. A., Gronlund, S. D., Clark, S. E., Carlson, C. A., Goodsell, C. A., Weatherford, D., & **Wixted, J. T.** (2017). ROCs in eyewitness identification: Instructions versus confidence ratings. *Applied Cognitive Psychology, 31*, 467–477. DOI: https://doi.org/10.1002/acp.3344

Miller, S. L., Zielaskowski, K., & **Plant, E. A.** (2012). The basis of shooter biases: Beyond cultural stereotypes. *Personality and Social Psychology Bulletin, 38*, 1358–1366. DOI: https://doi.org/10.1177/0146167212450516

Musolino, E. (2012). The influence of target gender on shooting behaviour: An examination of the role of automatic bias and controlled processes. Master's Thesis: Carlton University.

Nisbett, R. E., & **Wilson, T. D.** (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231–259. DOI: https://doi.org/10.1037/0033-295X.84.3.231

Osth, A. F., Bora, B., Dennis, S., & **Heathcote, A.** (2017). Diffusion vs. linear ballistic accumulation: Different models, different conclusions about the slope of the zROC in recognition memory. *Journal of Memory & Language, 96*, 36–61. DOI: https://doi.org/10.1016/j.jml.2017.04.003

Park, S. H., & **Glaser, J.** (2011). Implicit motivation to control prejudice and exposure to counterstereotypic instances reduce spontaneous discriminatory behavior. *Korean Journal of Social and Personality Psychology, 25*, 107–120.

Park, S. H., Glaser, J., & **Knowles, E. D.** (2008). Implicit motivation to control prejudice moderates the effect of cognitive depletion on unintented discrimination. *Social Cognition, 26*, 401–419. DOI: https://doi.org/10.1521/soco.2008.26.4.401

Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology, 81*, 181–192. DOI: https://doi.org/10.1037//0022-3514.81.2.181

Payne, B. K., Lambert, A. J., & **Jacoby, L. L.** (2002). Best laid plans: Effects of goals on accessibility bias and cognitive control in race-based misperceptions of weapons. *Journal of Experimental Social Psychology, 38*, 384–396. DOI: https://doi.org/10.1016/S0022-1031(02)00006-9

Pazzaglia, A. M., Dube, C., & **Rotello, C. M.** (2013). A critical comparison of discrete-state and continuous models of recognition memory: Implications for recognition and beyond. *Psychological Bulletin, 139*, 1173–1203. DOI: https://doi.org/10.1037/a0033044

Petrusic, W. M., & **Baranski, J. V.** (2003). Judging confidence influences decision processing in comparative judgments. *Psychonomic Bulletin & Review, 10*, 177–183. DOI: https://doi.org/10.3758/BF03196482

Plant, E. A., Goplen, J., & **Kunstman, J. W.** (2011). Selective responses to threat: The roles of race and gender in decisions to shoot. *Personality and Social Psychology Bulletin, 37*, 1274–1281. DOI: https://doi.org/10.1177/0146167211408617

Plant, E. A., & **Peruche, B. M.** (2005). The consequences of race for police officers' responses to criminal suspects. *Psychological Science, 16*, 180–183. DOI: https://doi.org/10.1111/j.0956-7976.2005.00800.x

Plant, E. A., Peruche, B. M., & **Butz, D. A.** (2005). Eliminating automatic racial bias: Making race non-diagnostic for responses to criminal suspects. *Journal of Experimental Social Psychology, 41*, 141–156. DOI: https://doi.org/10.1016/j.jesp.2004.07.004

Pleskac, T. J., Cesario, J., & **Johnson, D. J.** (in press). How race affects evidence accumulation during the decision to shoot. *Psychonomic Bulletin & Review*. DOI: https://doi.org/10.3758/s13423-017-1369-6

Ratcliff, R., McKoon, G., & **Tindall, M.** (1994). Empirical generality of data from recognition memory receiver operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 763–785. DOI: https://doi.org/10.1037/0278-7393.20.4.763

Ratcliff, R., Van Zandt, T., & **McKoon, G.** (1995). Process Dissociation, Single-Process Theories, and Recognition Memory. *Journal of Experimental Psychology: General, 124*, 352–374. DOI: https://doi.org/10.1037/0096-3445.124.4.352

Rotello, C. M. (2017). Signal detection theories of recognition memory. In: Wixted, J. T. (Ed.), *Learning and Memory: A Comprehensive Reference,* 2nd edition*, 4. Cognitive Psychology of Memory*. Elsevier. DOI: https://doi.org/10.1016/B978-0-12-809324-5.21044-4

Rotello, C. M., Heit, E., & **Dubé, C.** (2015). When more data steer us wrong: Replications with the wrong dependent measure perpetuate erroneous conclusions. *Psychonomic Bulletin & Review, 22*, 944–954. DOI: https://doi.org/10.3758/s13423-014-0759-2

Rotello, C. M., & **Macmillan, N. A.** (2008). Response bias in recognition memory. In: Benjamin, A. S., & Ross, B. H. (Eds.), *The Psychology of Learning and Motivation: Skill and Strategy in Memory Use*, *48*, 61–94. Academic Press, London. DOI: https://doi.org/10.1016/S0079-7421(07)48002-1

Rotello, C. M., Masson, M. E. J., & **Verde, M. F.** (2008). Type I error rates and power analyses for single-point sensitivity measures. *Perception & Psychophysics, 70*, 389–401. DOI: https://doi.org/10.3758/PP.70.2.389

Sadler, M. S., Correll, J., Park, B., & **Judd, C. M.** (2012). The world is not black and white: Racial bias in the decision to shoot in a multiethnic context. *Journal of Social Issues, 68*, 286–313. DOI: https://doi.org/10.1111/j.1540-4560.2012.01749.x

Senholzi, K. B., Depude, B. E., Correll, J., Banich, M. T., & **Ito, T. A.** (2015). Brain activation underlying threat detection to targets of different races. *Social Neuroscience, 10*, 651–662. DOI: https://doi.org/10.1080/17470919.2015.1091380

Sherman, J. W., Gawronski, B., Gonsalkorale, K., Hugenberg, K., Allen, T. J., & **Groom, C. J.** (2008). The self-regulation of automatic associations and behavioral impulses. *Psychological Review, 115*, 314–335. DOI: https://doi.org/10.1037/0033-295X.115.2.314

Sim, J. J., Correll, J., & **Sadler, M. S.** (2013). Understanding police and expert performance: When training attenuates (vs. exacerbates) stereotypic bias in the decision to shoot. *Personality and Social Psychology Bulletin, 39*, 291–304. DOI: https://doi.org/10.1177/0146167212473157

Simmons, J. P., Nelson, L. D., & **Simonsohn, U.** (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366. DOI: https://doi.org/10.1177/0956797611417632

Smith, S. M., Roster, C. A., Golden, L. L., & **Albaum, G. S.** (2016). A multi-group analysis of online survey respondent data quality: Comparing a regular USA consumer panel to MTurk samples. *Journal of Business Research, 69*, 3139–3148. DOI: https://doi.org/10.1016/j.jbusres.2015.12.002

Snodgrass, J. G., & **Corwin, J.** (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117*, 34–50. DOI: https://doi.org/10.1037/0096-3445.117.1.34

Swets, J. A. (1986a). Indices of Discrimination or Diagnostic Accuracy: Their ROCs and Implied Models. *Psychological Bulletin, 99*, 100–117. DOI: https://doi.org/10.1037/0033-2909.99.1.100

Swets, J. A. (1986b). Form of Empirical ROCs in Discrimination and Diagnostic Tasks: Implications for Theory and Measurement of Performance. *Psychological Bulletin, 99*, 181–198. DOI: https://doi.org/10.1037/0033-2909.99.2.181

Swets, J. A., Pickett, R. M., Whitehead, S. F., Getty, D. J., Schnur, J. A., Swets, J. B., & **Freeman, B. A.** (1979). Assessment of diagnostic technologies. *Science, 205*, 753–759. DOI: https://doi.org/10.1126/science.462188

Swets, J. A., Tanner, W. P., Jr., & **Birdsall, T. G.** (1961). Decision processes in perception. *Psychological Review, 68*, 301–340. DOI: https://doi.org/10.1037/h0040547

Taylor, A. (2011). The influence of target race on split-second shooting decisions in simulated scenarios: A Canadian perspective. Order No. AAINR83227, *Dissertation Abstracts International. Section B: The Sciences and Engineering*.

Tenhundfeld, N. L. (2015). The gun wielding bias embodiment effect under stress. Masters Thesis.

Van Ravenzwaaij, D., & **Oberauer, K.** (2009). How to use the diffusion model: Parameter recovery of three methods: EZ, fast-dm, and DMAT. *Journal of Mathematical Psychology, 53*, 463–473. DOI: https://doi.org/10.1016/j.jmp.2009.09.004

Van Zandt, T., & **Maldonado-Molina, M. M.** (2004). Response reversals in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 1147–1166. DOI: https://doi.org/10.1037/0278-7393.30.6.1147

Watt, J. J. (2010). Strategies for influencing implicit associations in shooting decisions: A test of longevity. *Doctoral Thesis*.

Witt, J. K., & **Brockmole, J. R.** (2012). Action alters object identification: Wielding a gun increases the bias to see guns. *Journal of Experimental Psychology: Human Perception and Performance, 38*, 1159–1167. DOI: https://doi.org/10.1037/a0027881

Xavier, R., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & **Müller, M.** (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12*, 77. DOI: https://doi.org/10.1186/1471-2105-12-77

Peer review comments
The author(s) of this paper chose the Open Review option, and the peer review comments are available at: http://doi.org/10.1525/collabra.171.pr

How to cite this article: Rotello, C. M., Kelly, I. J., & Heit, E. (2018). The Shape of ROC Curves in Shooter Tasks: Implications for Best Practices in Analysis. *Collabra: Psychology,* 4(1): 32. DOI: https://doi.org/10.1525/collabra.171

Senior Editor: Simine Vazire

Editor: Ed Vul

Submitted: 25 May 2018 **Accepted:** 09 August 2018 **Published:** 31 August 2018

UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Collabra: Psychology is a peer-reviewed open access journal published by University of California Press.

OPEN ACCESS